Tracking-Guided 4D Generation: Foundation-Tracker Motion Priors for 3D Model Animation

Supplementary Material

1. Implementation Details

For multi-view video diffusion training, we uniformly sample 16 frames per animation. We optimize with AdamW (learning rate 1e-6, weight decay 0.01) using a constant-with-warmup LR scheduler for 20 epochs with a batch size of 1024 (16 × 4 views × 16 frames) on 16 NVIDIA H100 GPUs. At inference, we use 25 sampling steps to obtain stable 3D object animations. For 4D reconstruction, the Hexplane resolution and feature dimension are set to [100, 100, 8] and 16, respectively. We perform motion reconstruction progressively for the first 750 iterations with a batch size of 64 (4 views × 16 frames), and then enable 4D-SDS and ARAP for an additional 250 iterations. The learning rates are 0.01 for Hex-planes and 1e-4 for the offset prediction layers. The loss weights $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6$ are set to 1.0, 0.1, 10, 100.0, 0.01 and 10.0 respectively. Training the multi-view diffusion model takes around 12-14 hours on 16 × 80GB H100 GPUs. Per-object 4D optimization takes 40 minutes on a single A100 GPU (around 20 minutes for motion reconstruction and around 20 minutes for 4D-SDS and ARAP).

2. Evalaution Metrics Details

Multi-view Video Generation Evaluation Metrics: Following VBench [1], we evaluate multi-view video generation with five complementary metrics that jointly capture appearance/semantic fidelity, motion quality, and aesthetics: 1) I2V: image-to-video consistency computed in the DINO feature space, quantifying appearance alignment between the conditioning image and the generated video; 2) Motion Smoothness (M.,Sm): temporal smoothness estimated by comparing the generated sequence with frames synthesized via interpolation, penalizing jitter and discontinuities across time; 3) Temporal Fidelity (T.,Fli): tability of scene content under viewpoint changes, measuring how well identities, geometry, and textures persist over the sequence; 4) Dynamic Score (Dy.,Sc): a continuous estimate of motion intensity derived from the same flow-based predictor, reflecting the richness and realism of dynamics without over- or under-motion; 5) Aesthetic Quality (Aest.,Q):perceptual appeal scored by the LAION Aesthetic Predictor, capturing overall visual quality of rendered frames.

4D Generation Evaluation Metrics: Our evaluation of 4D asset generation includes CLIP-based semantic alignment and a human user study. Following prior work [6, 7], we evaluate 4D-GS reconstructions with five CLIP-based similarity metrics computed in the CLIP embedding space, aggregating over views/frames as specified below: 1) CLIP-O(img): CLIP similarity between the rendered object image and the reference static object image; 2) CLIP-O(text): CLIP similarity between the rendered object image and the text prompt describing that object; 3) CLIP-F(img): frame-wise CLIP similarity between rendered video frames and the reference static object image; 4) CLIP-F(text): frame-wise CLIP similarity between rendered video frames and the text prompt; 5) CLIP-C: composite score aggregating object-/frame-level, image-/text-conditioned similarities.

Following prior work in Animate3D [2], we further conduct a user study evaluated along four different criteria: 1) Alignment Text (Align. Text). Measures how closely the generated 4D asset (appearance + motion) semantically matches the input prompt-covering identity, attributes (color/material/parts), and action semantics. Raters assign 1-5 Likert scores (higher is better). 2) Alignment 3D (Align. 3D). Assesses geometric and canonical look consistency between the generated 4D asset and the provided static reference, emphasizing shape fidelity, fine textures, and identity preservation across views/time. Scored 1-5. 3) Motion (Mot.). Evaluates perceived motion quality, including physical plausibility, temporal smoothness, and continuity of articulations, while penalizing jitter, ghosting, or flicker. Scored 1-5. 4) Appearance (Appr.). Rates per-frame visual fidelity and aesthetics-sharpness, texture richness, material/lighting realism-and penalizes artifacts such as blur, seams, or ringing. Scored 1-5.

3. Additional Baseline

We provide additional comparisons against the EG4D [5] baseline in Tab. 1 and Tab. 2 for both multi-view video generation and 4D generation, using the same evaluation metrics as in the main manuscript. As shown in Tab. 1, our method surpasses EG4D across all video metrics, with especially pronounced gains in I2V (+0.23) and Aest. Q (+0.13), reflecting notably improved cross-view consistency and perceptual quality. The uniform improvements on M. Sm, T. Fli, and Dy. Sc further indicates that our model enhances motion smoothness, temporal stability and dynamic coherence. Tab. 2 reports 4D generation results on datasets Sketchfab28 and Animate3D. Our approach delivers clear and consistent improvements over EG4D on CLIP-O and CLIP-F under both image- and text-conditioned evaluations (e.g., +0.13-0.18 on CLIP-O/F(img) and +0.01-0.05

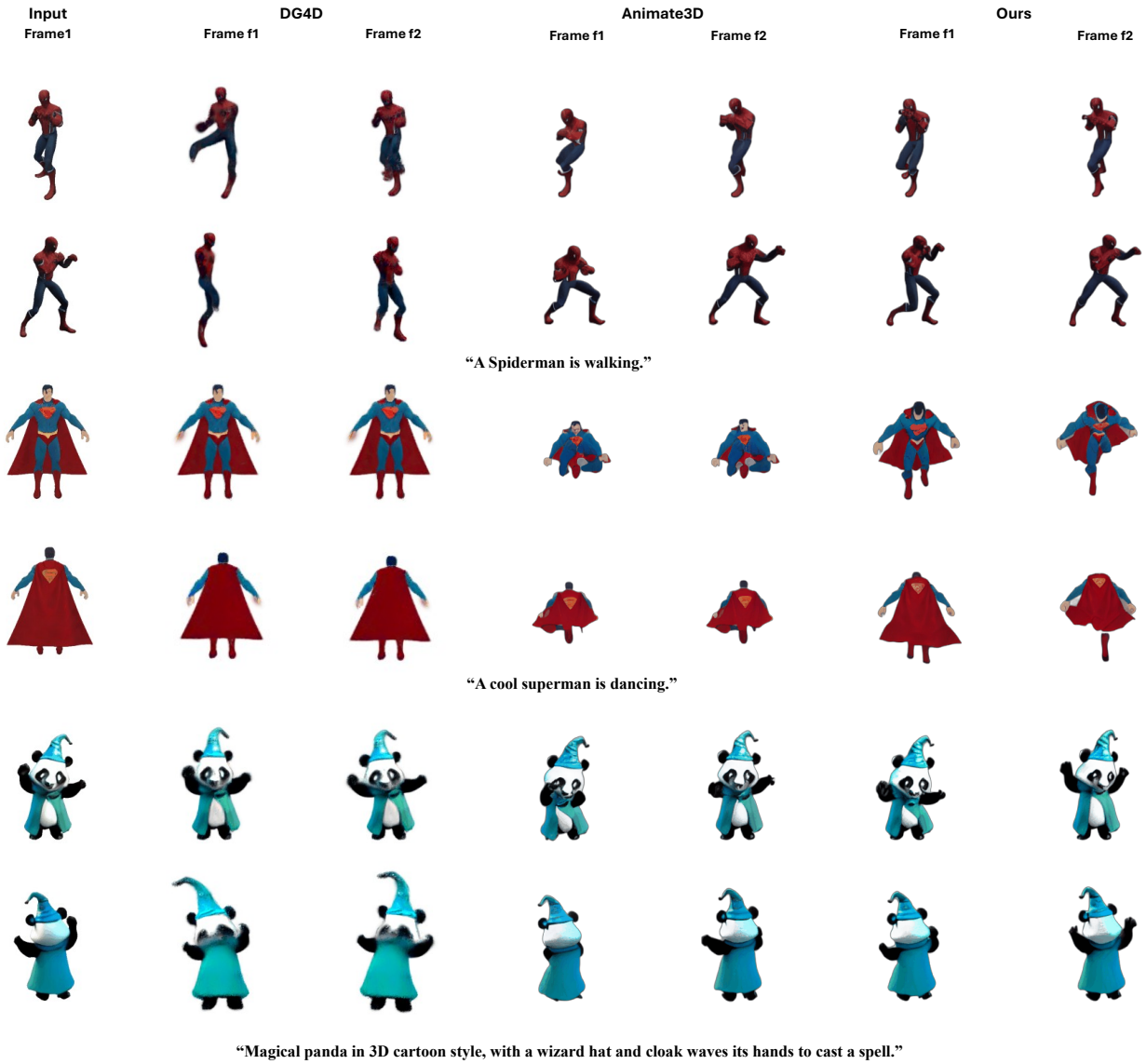

Figure 1. Video Generation qualitative comparison on Animate3D dataset. Best viewed by zooming in.

	I2V ↑	M. Sm ↑	T. Fli ↑	Dy. Sc ↑	Aest. Q ↑
EG4D	0.712	0.971	0.970	0.923	0.407
Ours	**0.933**	**0.992**	**0.991**	**1.356**	**0.470**

	I2V ↑	M. Sm ↑	T. Fli ↑	Dy. Sc ↑	Aest. Q ↑
EG4D	0.716	0.967	0.961	0.712	0.374
Ours	**0.945**	**0.993**	**0.992**	**0.778**	**0.506**

Table 1. Video Generation quantitative comparison on datasets: Diffusion4D (left); Animate3D (right).

	CLIP-O(img) ↑	CLIP-O(text) ↑	CLIP-F(img) ↑	CLIP-F(text) ↑	CLIP-C ↑
EG4D	0.7548	0.2374	0.7457	0.2564	0.9210
Ours	**0.8884**	**0.2664**	**0.8955**	**0.2664**	**0.9819**

	CLIP-O(img) ↑	CLIP-O(text) ↑	CLIP-F(img) ↑	CLIP-F(text) ↑	CLIP-C ↑
EG4D	0.7990	0.2598	0.7675	0.2548	0.9411
Ours	**0.9428**	**0.2980**	**0.9447**	**0.3005**	**0.9844**

Table 2. 4D Generation quantitative comparison on datasets: Sketchfab28 (left); Animate3D (right).

on CLIP-O/F(text)). Moreover, the higher CLIP-C scores across both datasets confirm stronger semantic alignment and cross-modal consistency in the generated 4D content. Overall, our approach consistently surpasses EG4D across both benchmarks, demonstrating clear gains in video quality and 4D reconstruction fidelity.

4. More Comparison Visualizations

We provide all the qualitative visualization results for both video generation and 4D generation in Figs. 1, 2 and 3, covering diverse objects, viewpoints, and motion regimes. Please note that for some DG4D examples, the rendered view is not perfectly aligned with the input view; for in-

Figure 2. 4D Generation qualitative comparison on Sketchfab28 dataset. Best viewed by zooming in.

stance, the Spider-Man result in the second row of Fig. 1. This mismatch arises from object position movement introduced by the DG4D codebase during rendering.

5. More Visualizations of Ablation Studies

We provide additional ablation study examples with clearer visual comparisons in Fig. 4, which more explicitly demonstrate the contributions of Corrs. Loss and Di. Feat.

6. More Details on New Evaluation Dataset

This section details the construction of our new evaluation dataset, *Sketchfab28* dataset. We processed 28 object meshes by baking realistic lighting directly into their vertex colors. To achieve consistent and high-quality illumination, we used a standard four-point lighting setup in Blender, consisting of a key light, fill light, back light, and background light. The key light provided the primary illumination to define the object's shape, while the fill light softened shadows, the back light enhanced depth by separating the object from the background, and the background light ensured even lighting of the scene. During baking, both direct and indirect light contributions were included, with all major shading components enabled, such as diffuse, glossy, transmission, and emission. This allowed the baked vertex colors to capture realistic shading and reflectance effects without relying on external texture maps. The baked meshes were then converted into 3D Gaussian splats, which serve as the input representation for subsequent stages of our pipeline. The *Sketchfab28* dataset can be found on the website[1].

7. Limitation Discussion

Highly Complex 4D Motions: Unlike 2D video generation, data collection and annotation for 4D generation are far more challenging. Unfortunately, existing 4D datasets do not include highly complex 4D object motions. Since

[1] https://github.com/dachengxiaocheng/Sketchfab28-Dataset

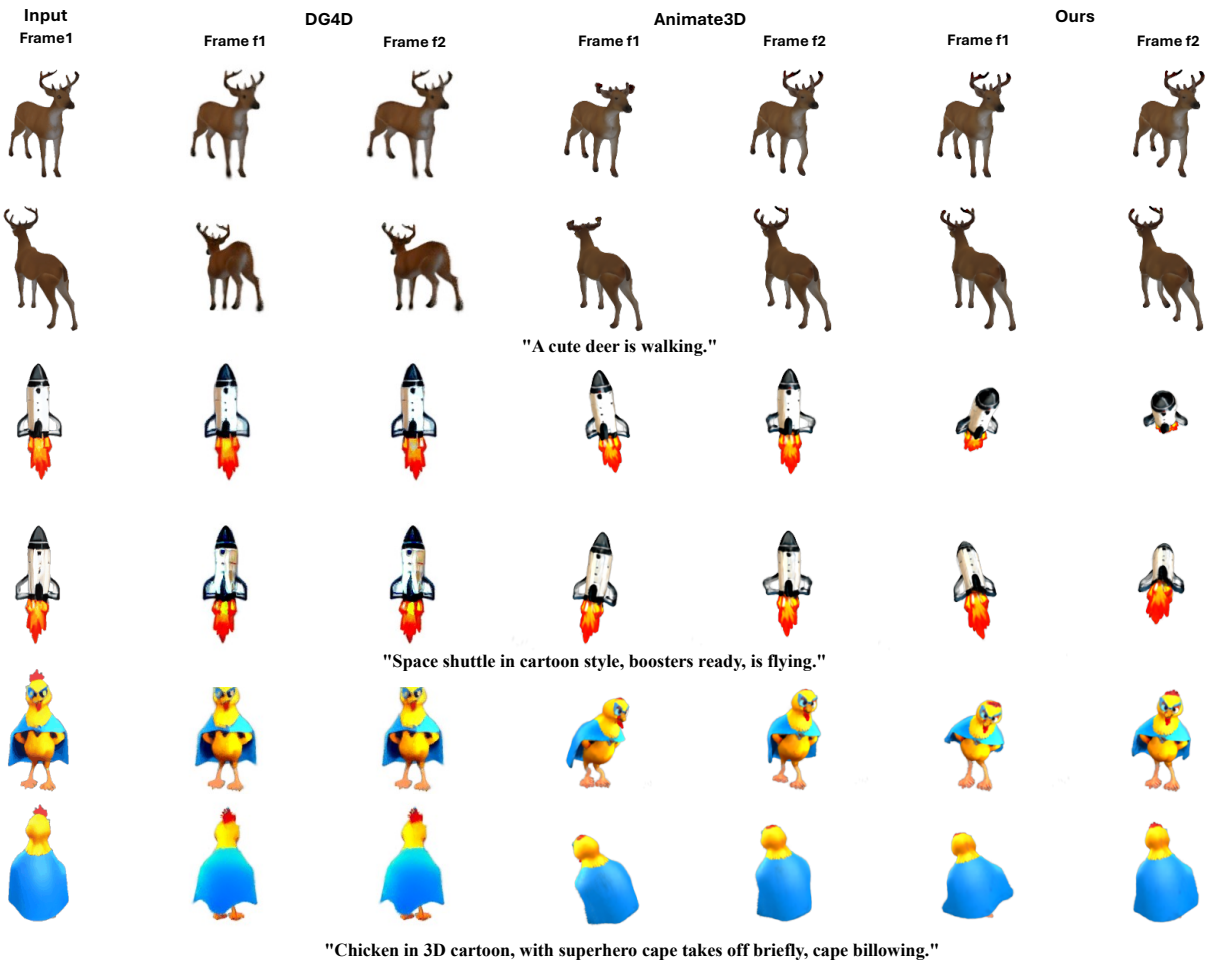

Figure 3. 4D Generation qualitative comparison on Animate3D dataset. Best viewed by zooming in.

generative models are highly data-driven, a model trained on such data cannot predict highly complex 4D motion.

CoTracker3's Capability: The feature-level temporal consistency constraint does depend on the accuracy of the chosen tracker. As a foundation tracking model, CoTracker3 [3] leverages joint multi-point reasoning on large-scale data and a pseudo-labeling training protocol, achieving strong data-scaling benefits that improve accuracy in fast motion, deformations and occlusion-heavy scenarios. We do not optimize or customize CoTracker3 features in our pipeline, we simply leverage a foundation tracker to provide pixel correspondences. Therefore, our framework is general and compatible with other high-performing trackers, which ensures that they can likewise yield performance gains. However, under extreme occlusion, CoTracker can lose track of the target. In such cases, introducing incorrect tracking constraints may lead to degraded model performance.

8. User Study Template

As shown in Fig. 5, we present the interface used in our user study. The survey evaluates the 4D objects from Sketchfab28 and Animate3D datasets. For each generated 4D object, participants provide 1-5 Likert ratings on three criteria: alignment with the given static object and text prompt, appearance fidelity, and motion quality.

9. Video Demos

We provide video demos on the website[2] for i) generated 4D assets and ii) multi-view point tracking, enabling synchronized multi-view comparisons. For the generated 4D assets, the early test cases, being more challenging, demonstrate that our approach significantly outperforms both baselines. In the later, comparatively simpler cases, our method still slightly surpasses Animate3D [2] and clearly outperforms DG4D [4].

[2] https://github.com/dachengxiaocheng/Video-Demos-for-Track4DGen

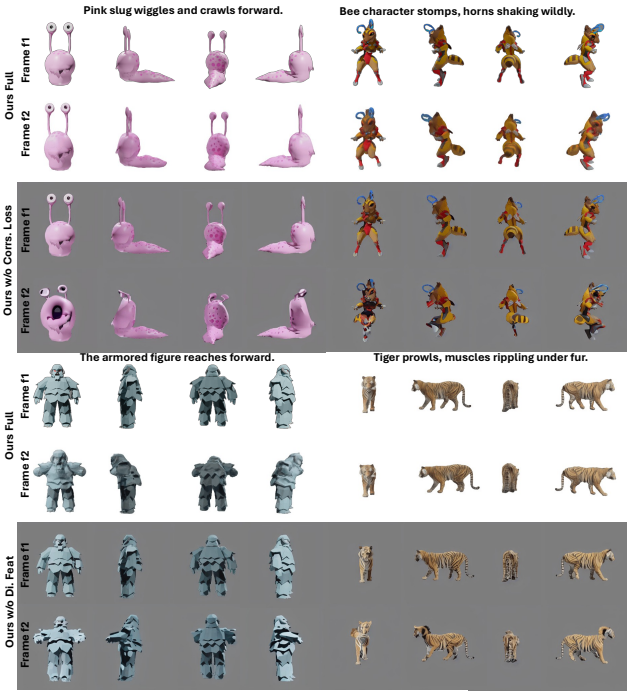

Figure 4. Ablation Study Visual Examples (Zooming in).

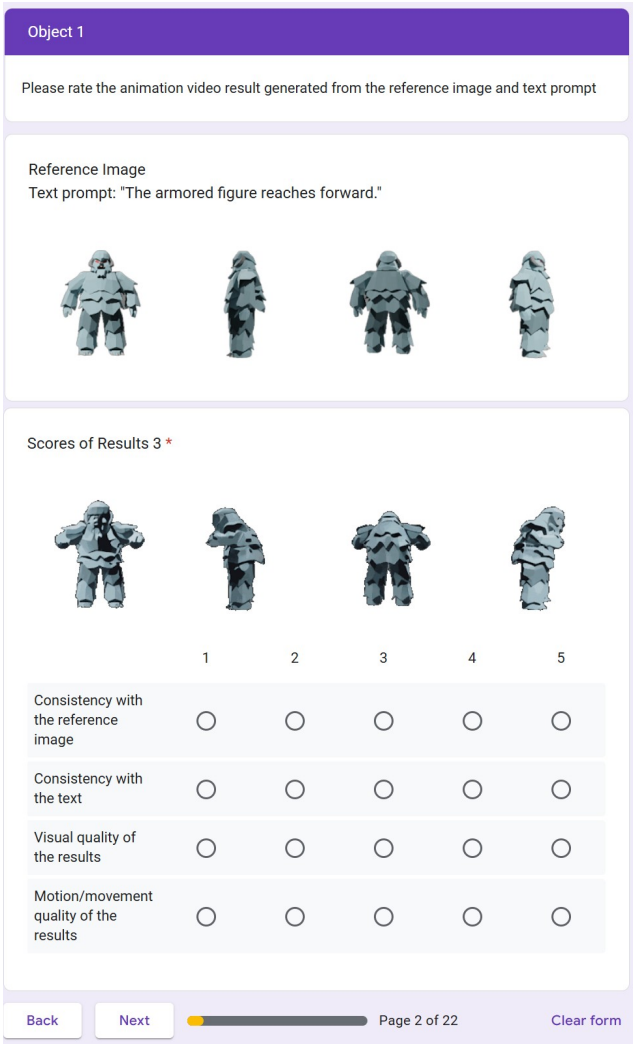

Figure 5. The UI of our user study.

References

[1] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 1

[2] Yanqin Jiang, Chaohui Yu, Chenjie Cao, Fan Wang, Weiming Hu, and Jin Gao. Animate3d: Animating any 3d model with multi-view video diffusion. *Advances in Neural Information Processing Systems*, 37:125879–125906, 2024. 1, 4

[3] Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6013–6022, 2025. 4

[4] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023. 4

[5] Qi Sun, Zhiyang Guo, Ziyu Wan, Jing Nathan Yan, Shengming Yin, Wengang Zhou, Jing Liao, and Houqiang Li. Eg4d: Explicit generation of 4d object without score distillation. In *International Conference on Learning Representations (ICLR)*, 2025. 1

[6] Cong Wan, Xiangyang Luo, Hao Luo, Zijian Cai, Yiren Song, Yunlong Zhao, Yifan Bai, Fan Wang, Yuhang He, and Yihong Gong. Grid: Omni visual generation. *arXiv preprint arXiv:2412.10718*, 2024. 1

[7] Haiyu Zhang, Xinyuan Chen, Yaohui Wang, Xihui Liu, Yunhong Wang, and Yu Qiao. 4diffusion: Multi-view video diffusion model for 4d generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1